Characterizing Structural Regularities of Labeled Data in Overparameterized Models

Ziheng Jiang*[1 2] Chiyuan Zhang*[3] Kunal Talwar[4] Michael C. Mozer[3]

[1]University of Washington [2]OctoML [3]Google [4]Apple

*: equal contribution

ICML International Conference On Machine Learning

A Binary Chairs vs Non-Chairs Problem

a few same-class neighbors (weak regularity)

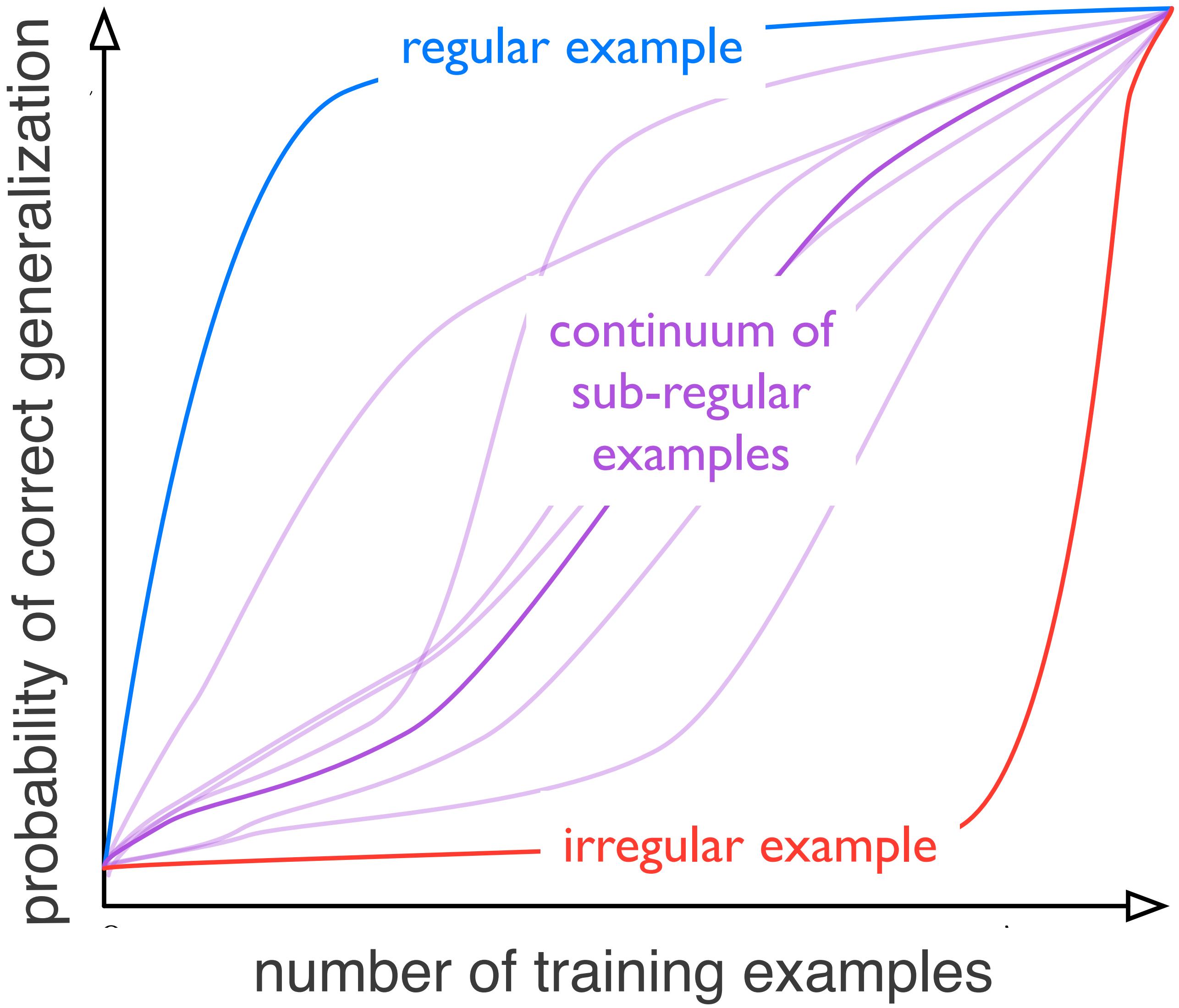
no neighbors of same class (irregular example)

many same-class neighbors (strong regularity)

If an example is held out from training, will a net generalize correctly?

A Continuum of Regularities

If we use n training examples to train the model, *the probability of correct generalization* for a specific instance will behave differently depends on the *structural regularities of the training data*.



The Consistency Profile and The C-score

Formalization

Consistency Profile

$$C_{\mathscr{P},n}(x,y) = \mathbb{E}_{D \overset{n}{\sim} \mathscr{P}}[\mathbb{P}(f(x; D \backslash \{(x,y)\}) = y]$$

data distribution

training set size

Formalization

Consistency Profile

$$C_{\mathscr{P},n}(x,y) = \mathbb{E}_{D \overset{n}{\sim} \mathscr{P}}[\mathbb{P}(f(x; D \backslash \{(x,y)\}) = y]$$

data distribution

training set size

Empirical Consistency Profile

$$\hat{C}_{\hat{\mathscr{D}},n}(x,y) = \mathbb{E}^{r}_{D \overset{n}{\sim} \hat{\mathscr{D}}}[\mathbb{P}(f(x; D \backslash \{(x,y)\}) = y]$$

data set

Formalization

Consistency Profile

$$C_{\mathcal{P},n}(x,y) = \mathbb{E}_{D \overset{n}{\sim} \mathcal{P}}[\mathbb{P}(f(x; D \backslash \{(x,y)\}) = y]$$

data distribution

training set size

Empirical Consistency Profile

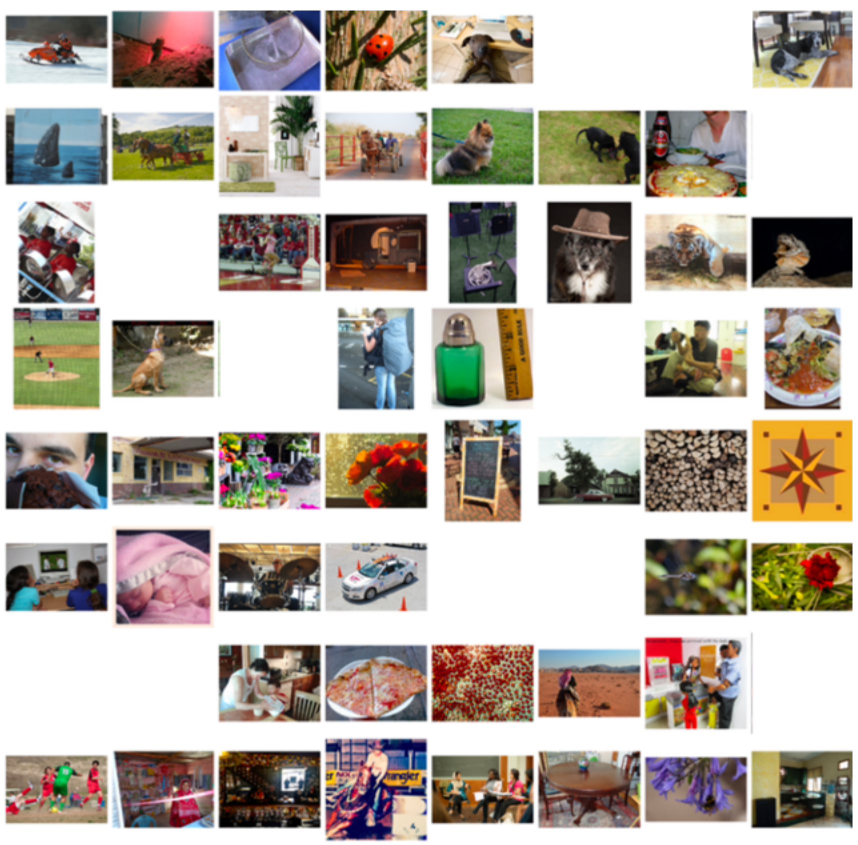
$$\hat{C}_{\hat{\mathcal{D}},n}(x,y) = \hat{\mathbb{E}}^{r}_{D \overset{n}{\sim} \hat{\mathcal{D}}}[\mathbb{P}(f(x; D \backslash \{(x,y)\}) = y]$$

data set

C Score

$$\hat{C}_{\hat{\mathcal{D}}}(x,y) = \mathbb{E}_{n}[\hat{C}_{\hat{\mathcal{D}},n}(x,y)]$$

Experiments for Empirical Estimation



Random subset of size n

Test on held out examples

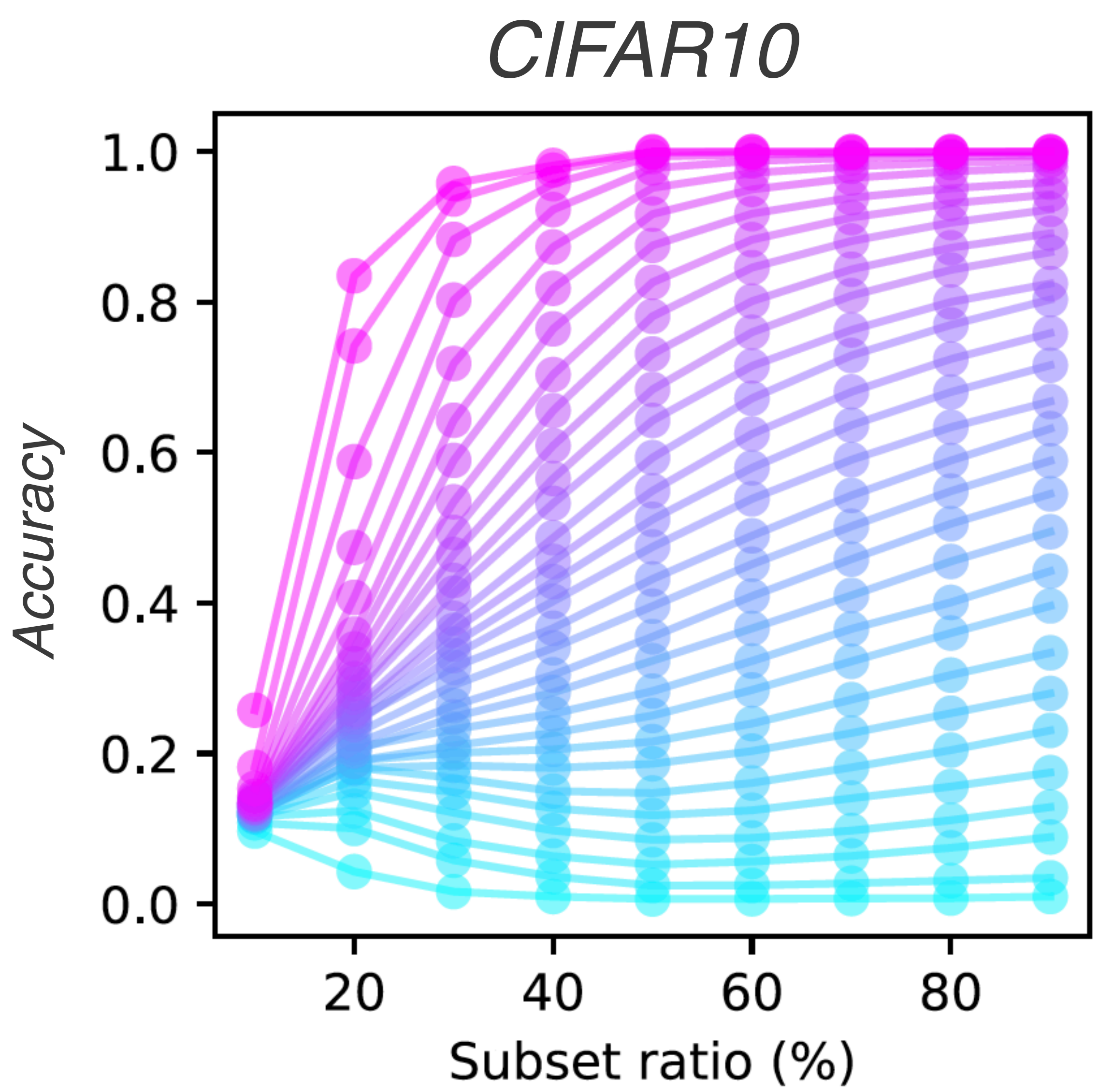
Training Pipeline → $\hat{f}$

data-augmentation, regularization,
stable initialization, SoTA activation function,
fancy lr schedule, momentum, preconditioning,
label smoothing, loss tempering, unsupervised aux loss

Repeat 2,000 Times

Empirical Consistency Profiles on CIFAR10

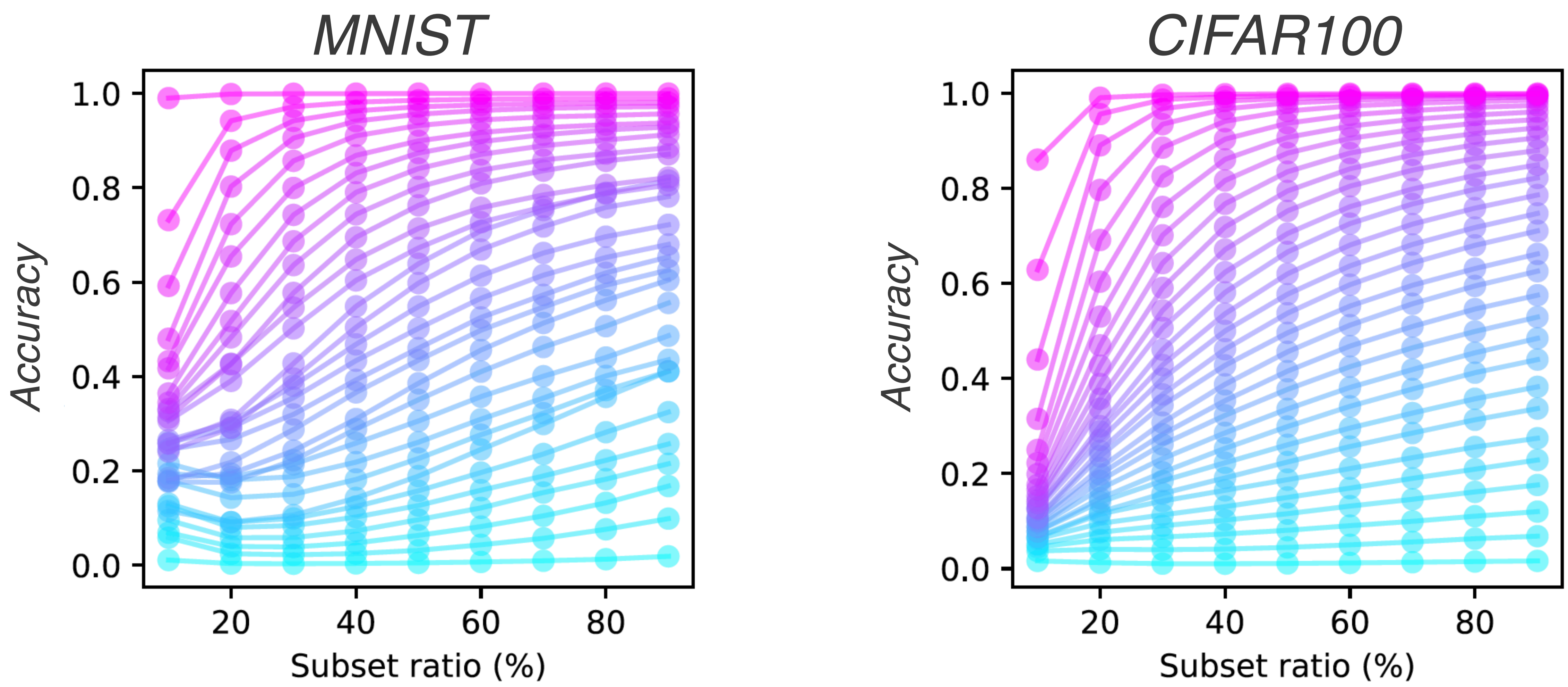
Each curve is a group of individual instances with similar mean accuracy.



As the subset ratio grows:

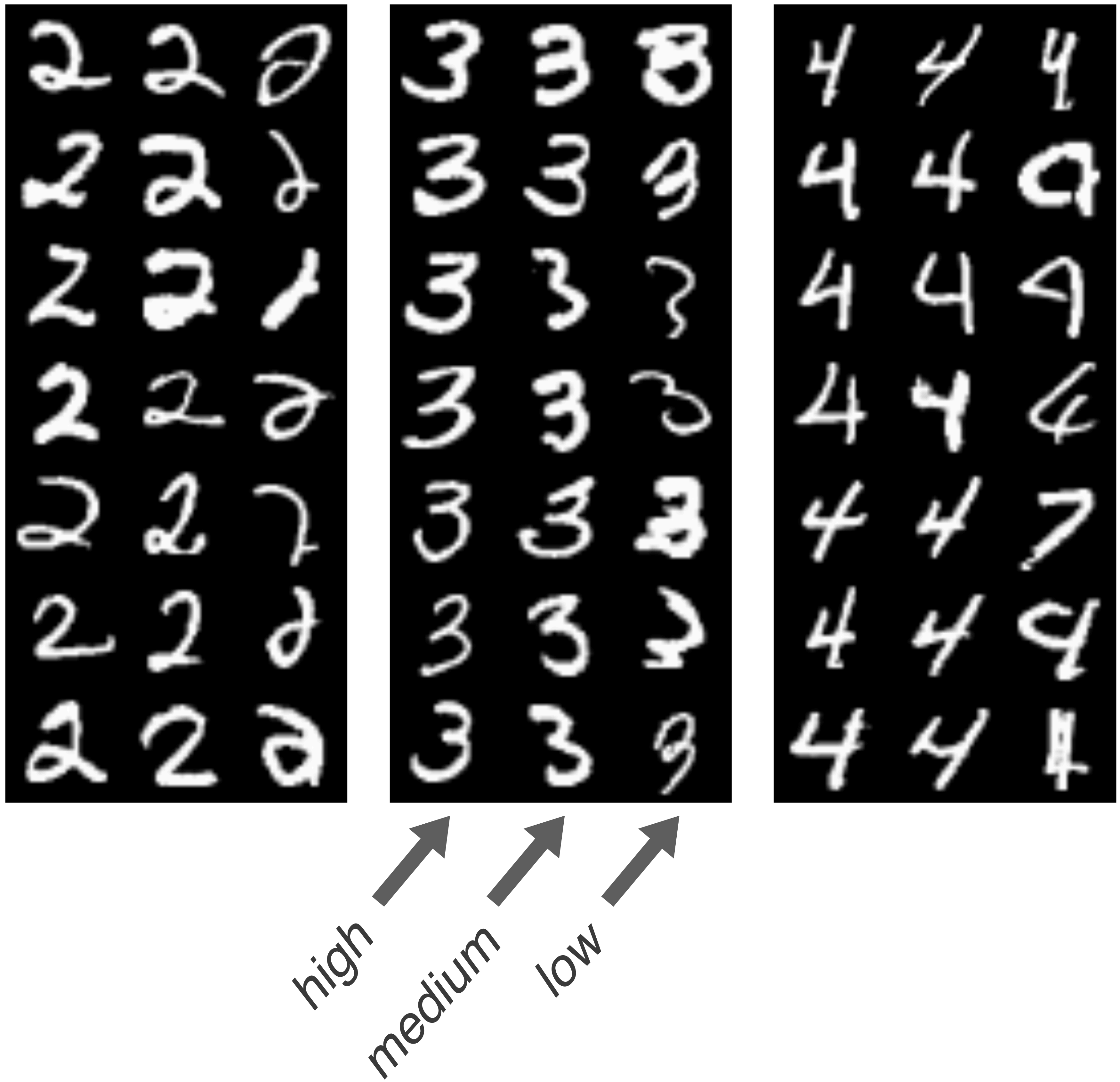
- the *top-ranked examples* can be classified correctly easily.
- the *bottom-ranked examples* have persistently low probability of correct classification.

Empirical Consistency Profiles on MNIST and CIFAR100

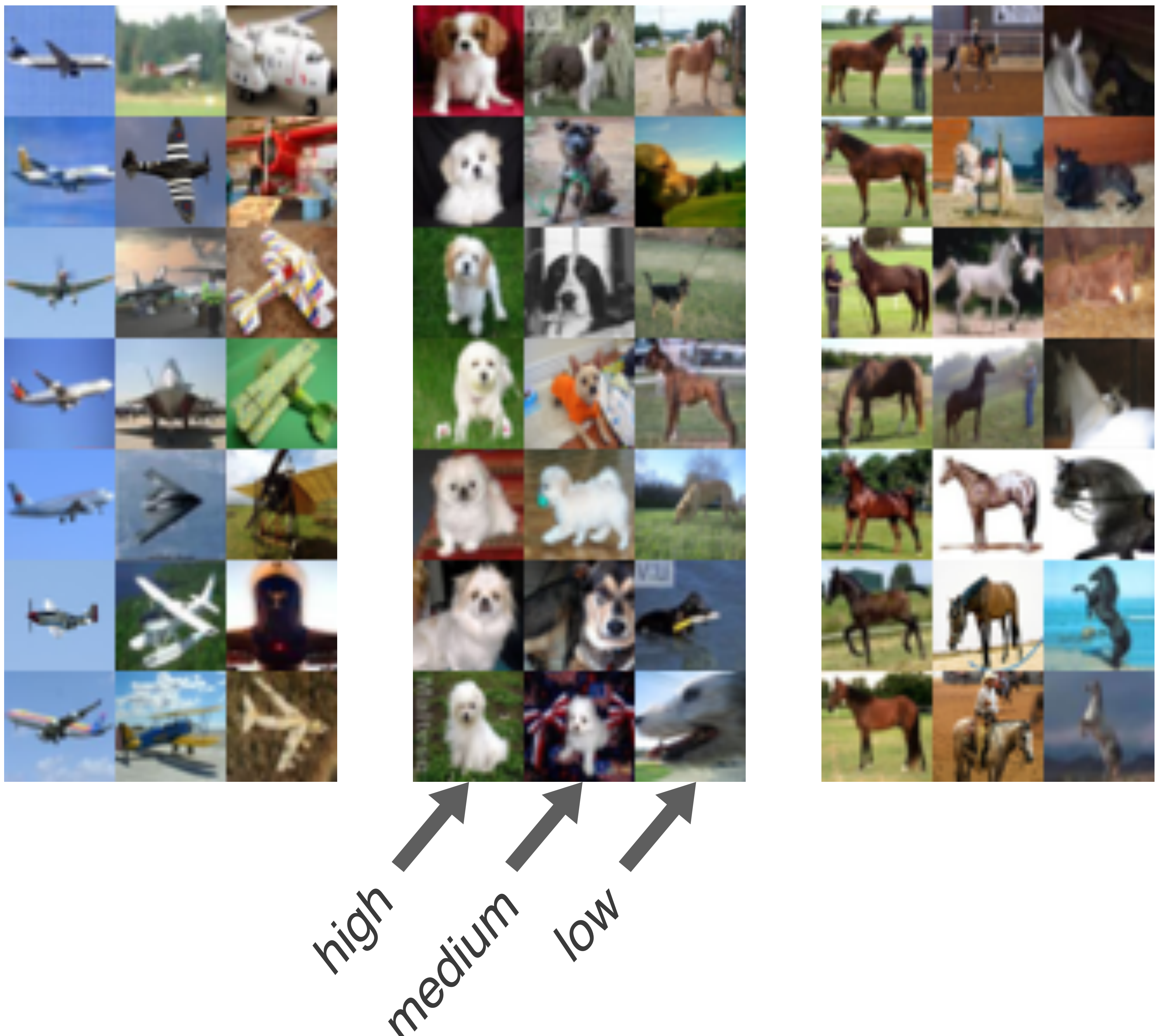


We observe similar results on MNIST and CIFAR100.

MNIST: Visualization of Examples Ranked by C-score

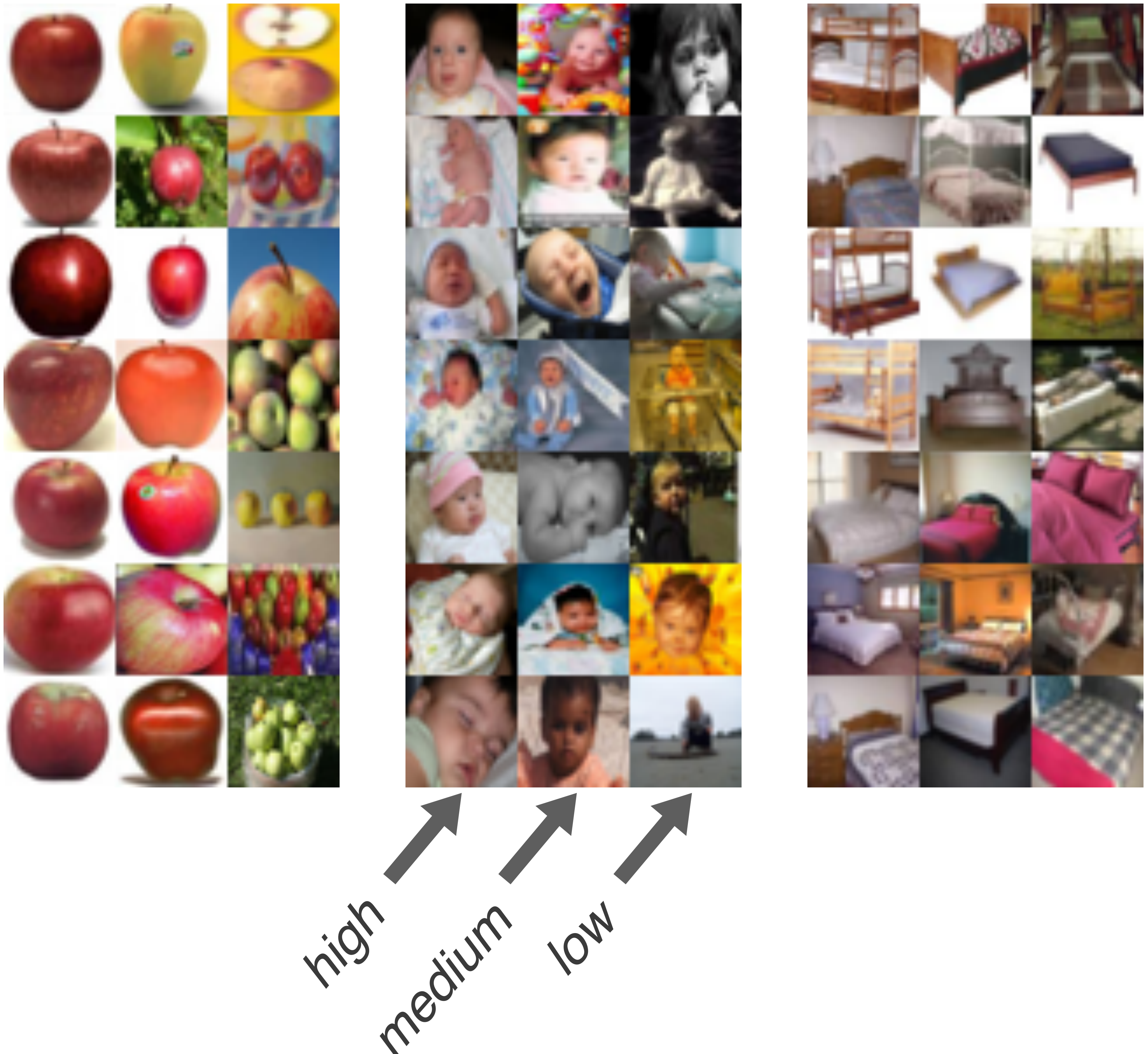


CIFAR10: Visualization of Examples Ranked by C-score

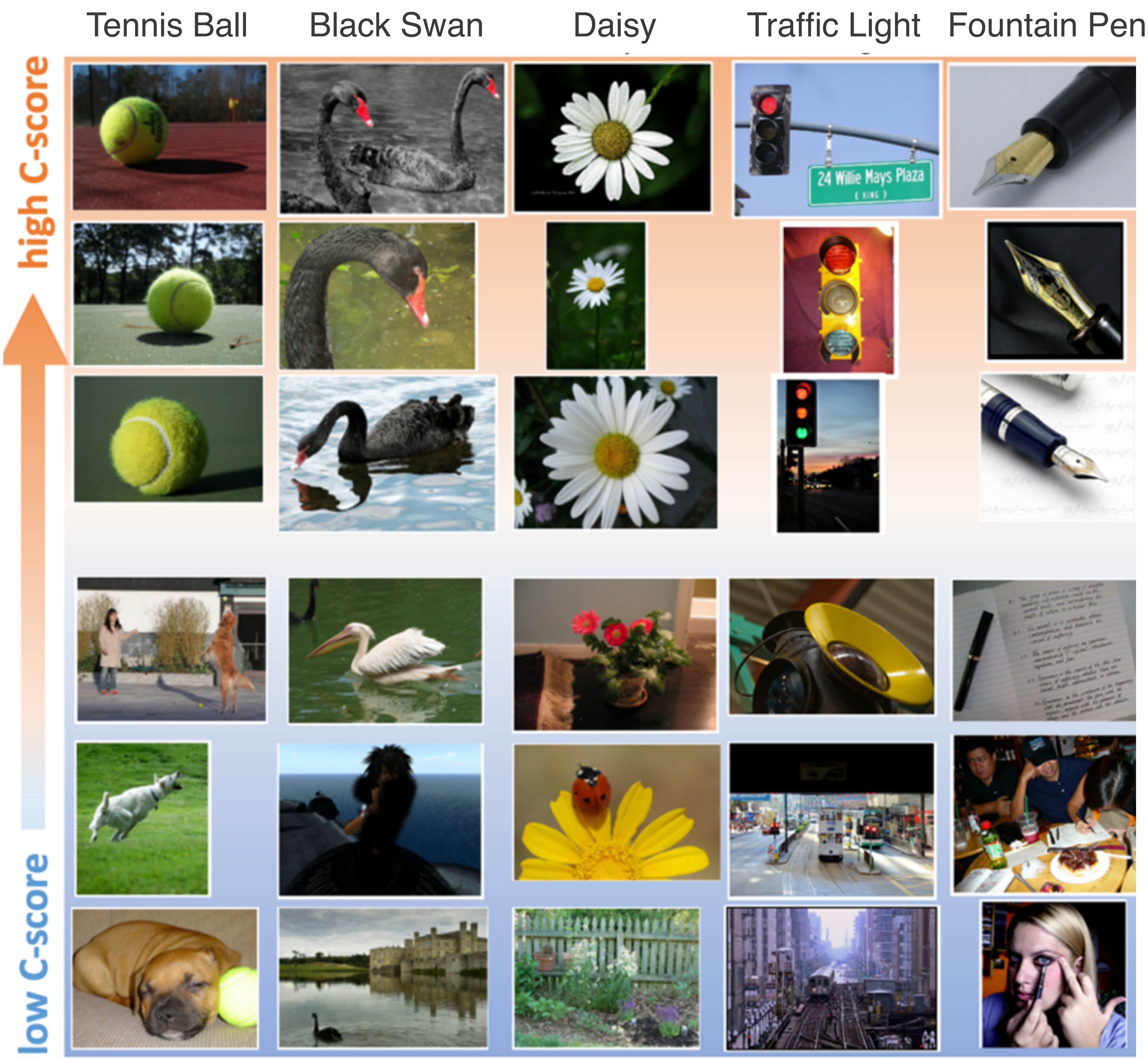


ICML
International Conference
On Machine Learning

CIFAR-100: Visualization of Examples Ranked by C-score



ImageNet: Visualization of Examples Ranked by C-score

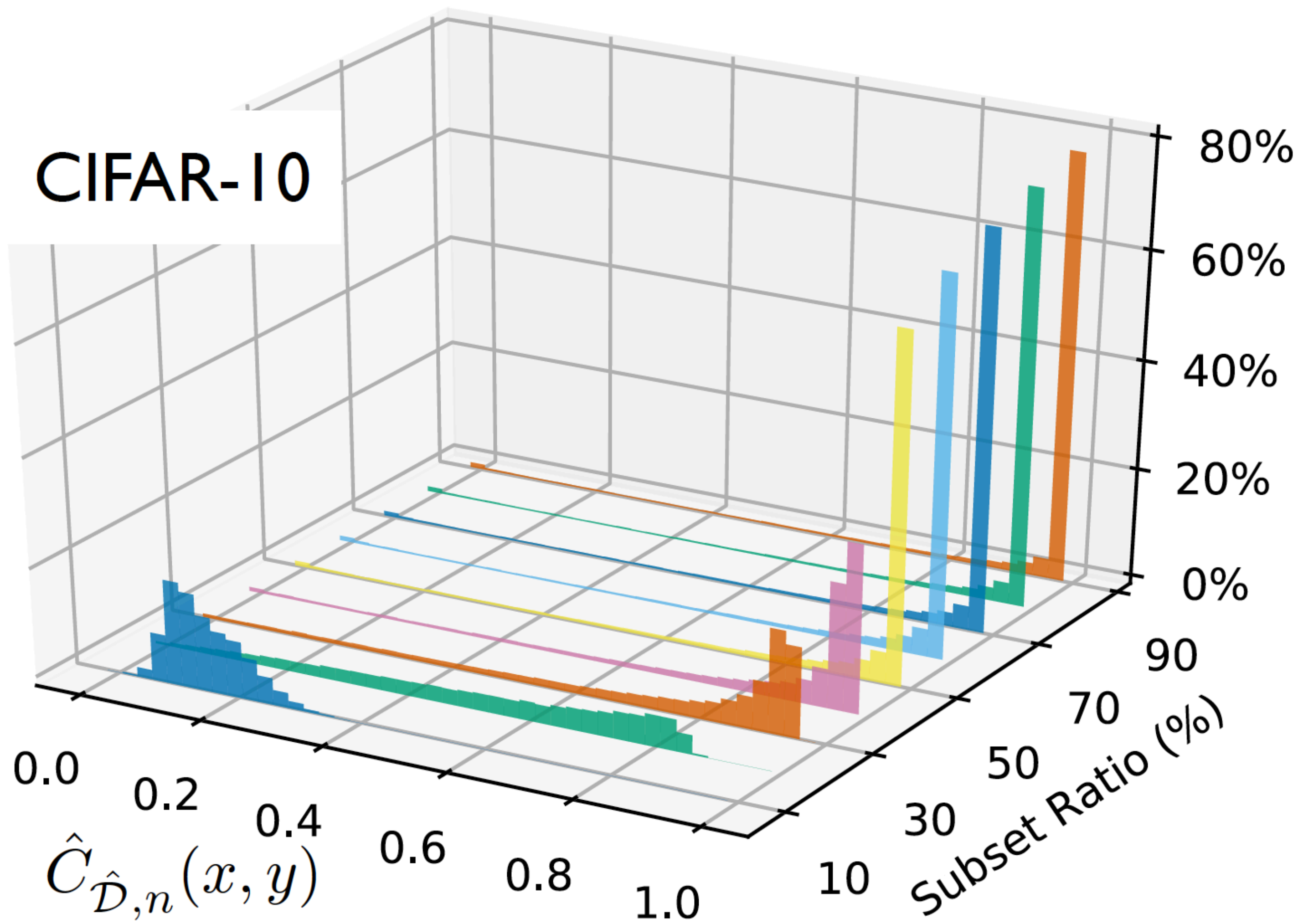


The Structural Regularities of Common Image Data Sets

ICML
International Conference
On Machine Learning

Floor and Ceiling Effects in The Empirical Consistency Profile

The distribution of the empirical consistency profile on CIFAR10

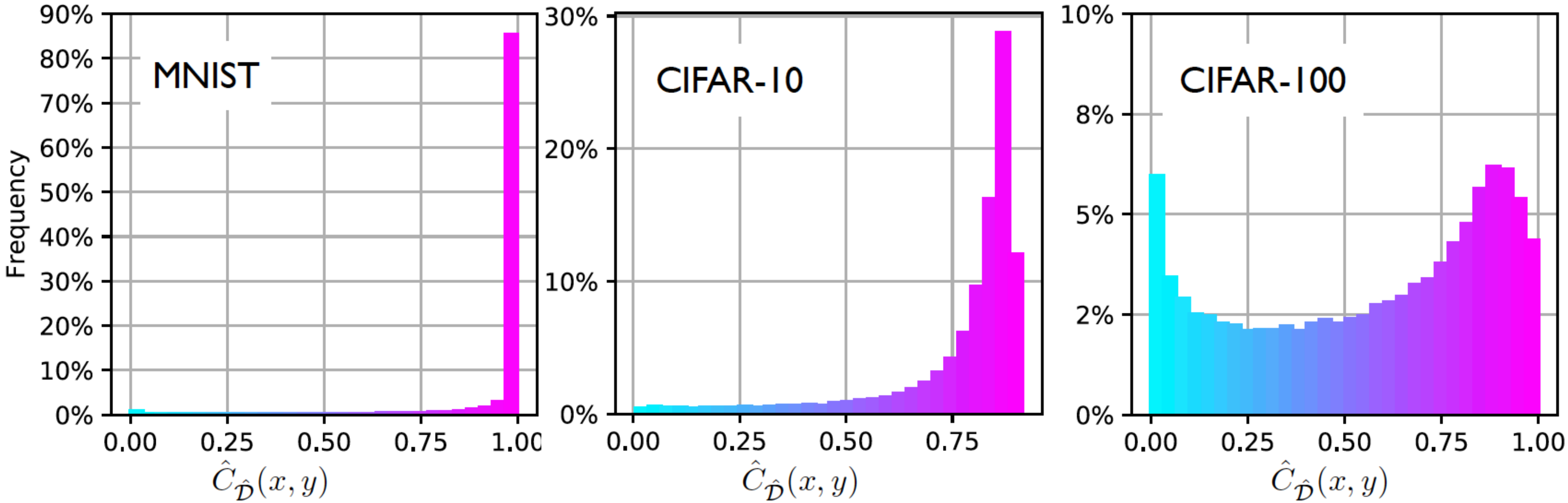


Depending on the subset ratio, instances may be *concentrated near the floor or ceiling*, making them difficult to distinguish

By taking an expectation over the subset ratio, the C-score is less susceptible to floor and ceiling effects.

ICML
International Conference
On Machine Learning

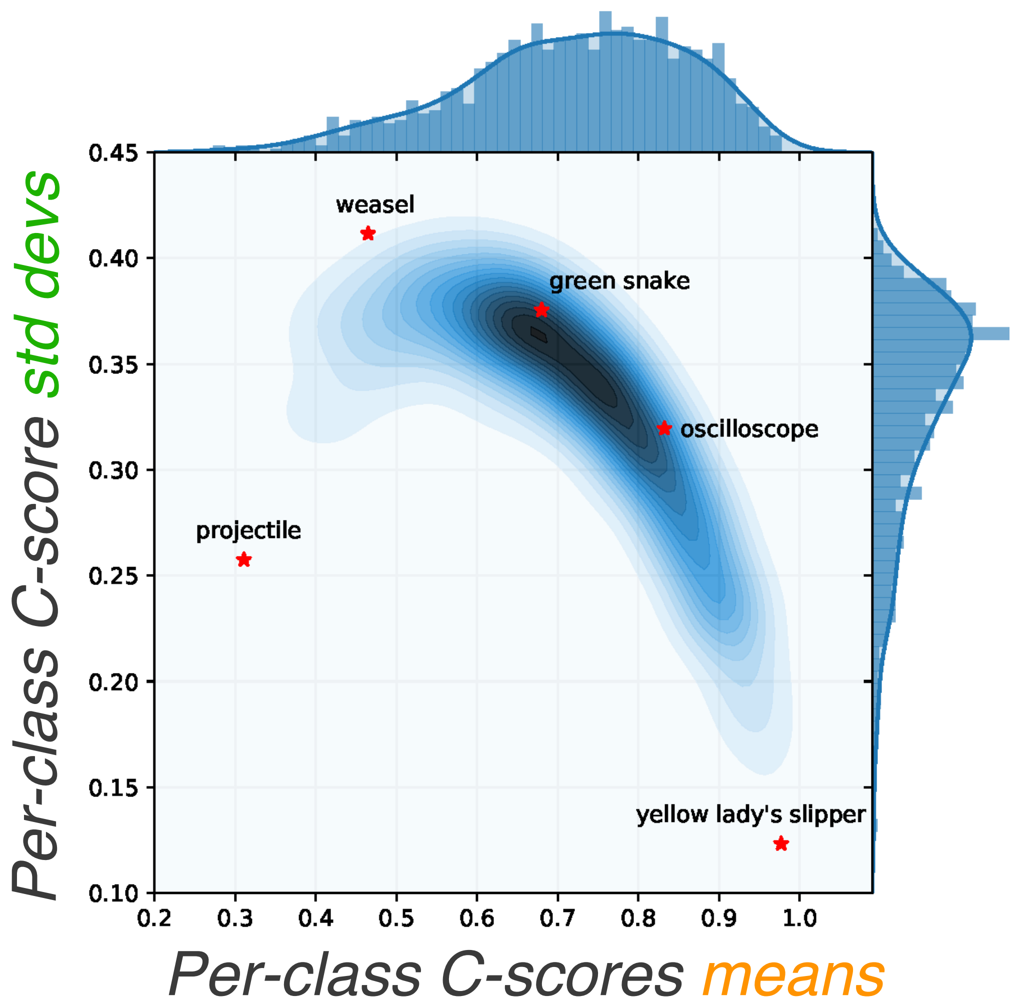
Histogram of The Integrated C-score



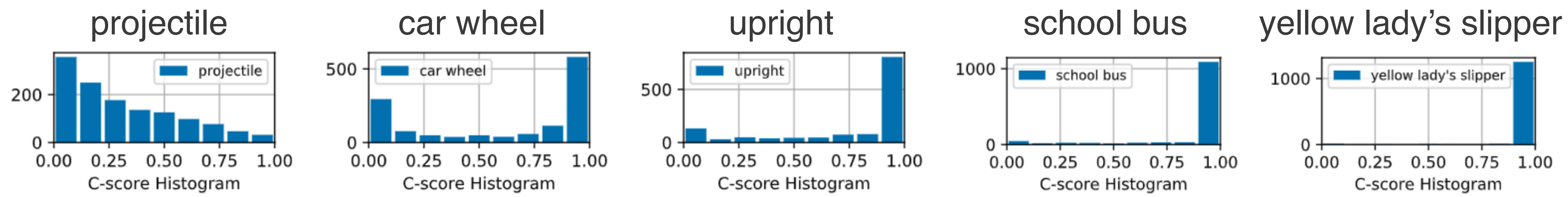
- The distribution is more uniformly spread than for specific subset ratios.
- The distribution reflects the structural regularities of data set.

The Structural Regularities of ImageNet

We compute the mean and standard deviation of the C-scores of all the examples for each class.



- The mean C-scores indicates the relative *difficulty* of classes.
- The standard deviation indicates the *diversity* of examples within each class.



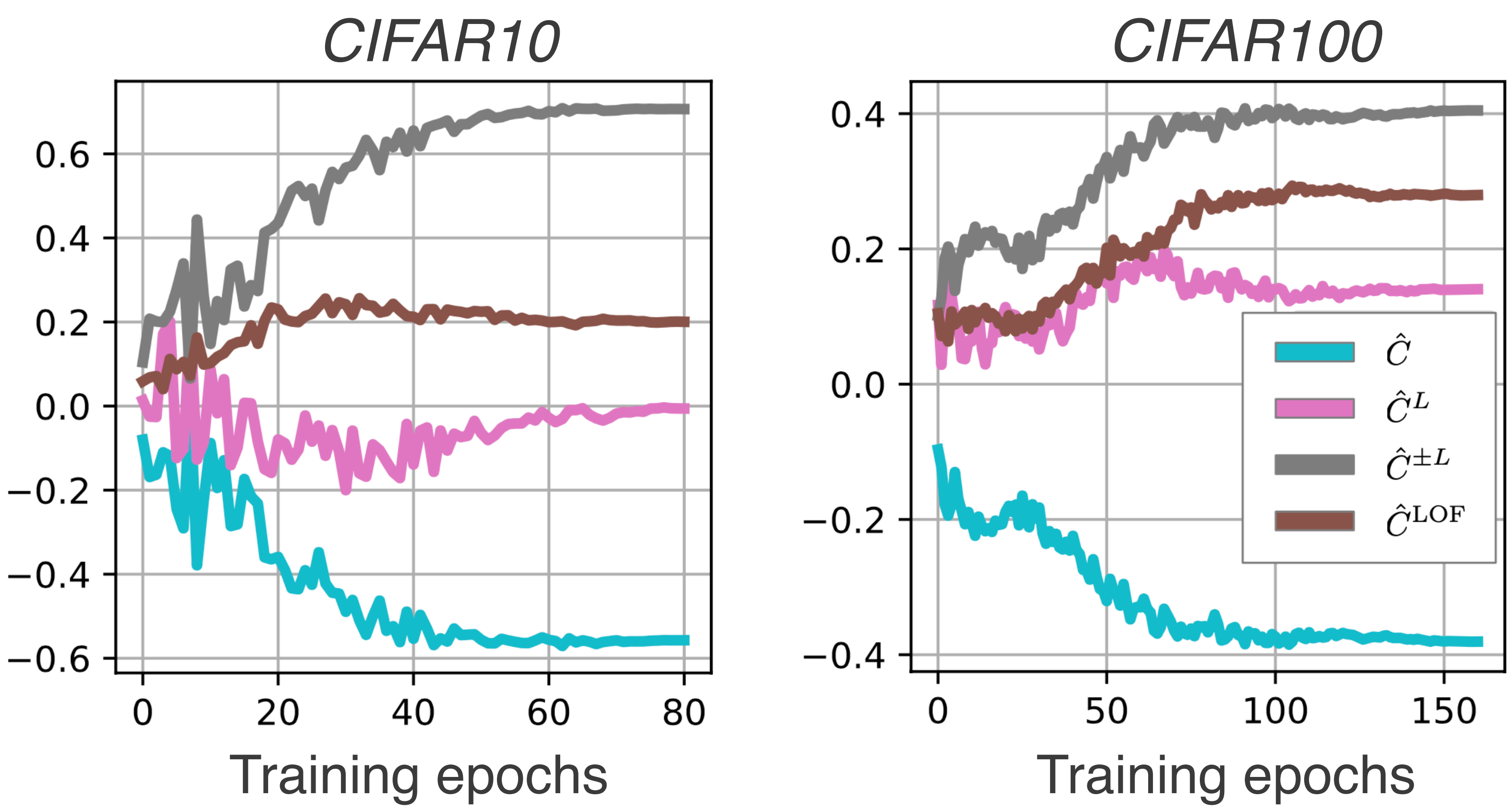
Efficient C-score Proxies

Pairwise Distance Based Proxies

We study four pairwise distance based proxies:

- $\hat{C}^{\pm L}(x, y)$: based on relative local density of all class labels
- $\hat{C}^{L}(x, y)$: based on relative local density of same-class examples
- $\hat{C}(x)$: based on relative local density, ignoring labels
- $\hat{C}^{LOF}(x)$: based on the LOF (local outlier factor) algorithm

Spearman rank correlation between C-score and distance based proxies



Conclusion:

The rankings are very sensitive to the underlying distance metrics

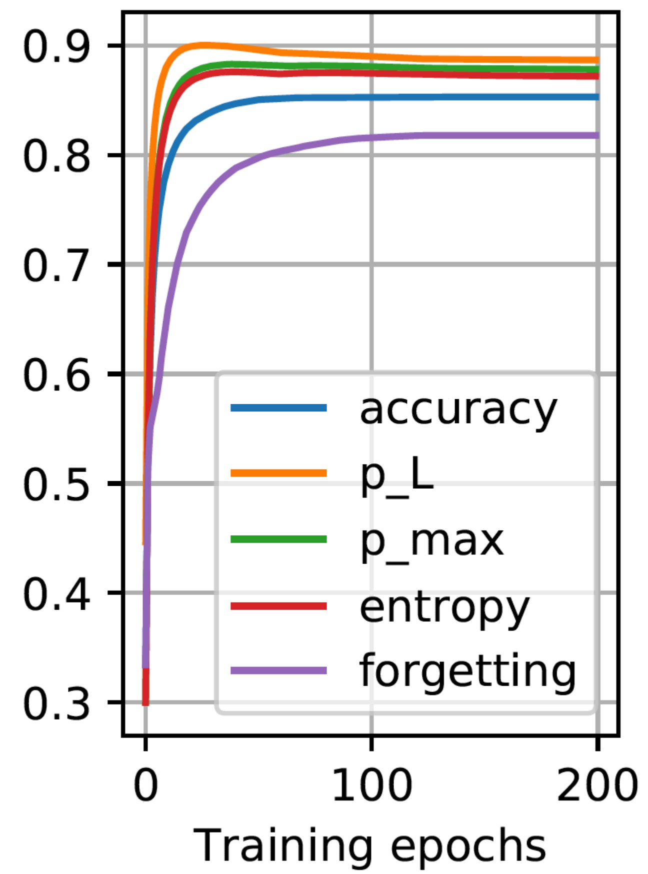
ICML
International Conference
On Machine Learning

Learning Speed Based Proxies

We study five learning speed based proxies:

- *accuracy*: based on 0-1 correctness
- p_L: based on softmax confidence on the correct class
- p_{max}: based on max softmax confidence across all classes
- *entropy*: based on negative entropy of softmax confidences
- *forgetting*: based on the forgetting event

Spearman rank correlation between C-score and learning speed based proxies on CIFAR-10.



Conclusion:

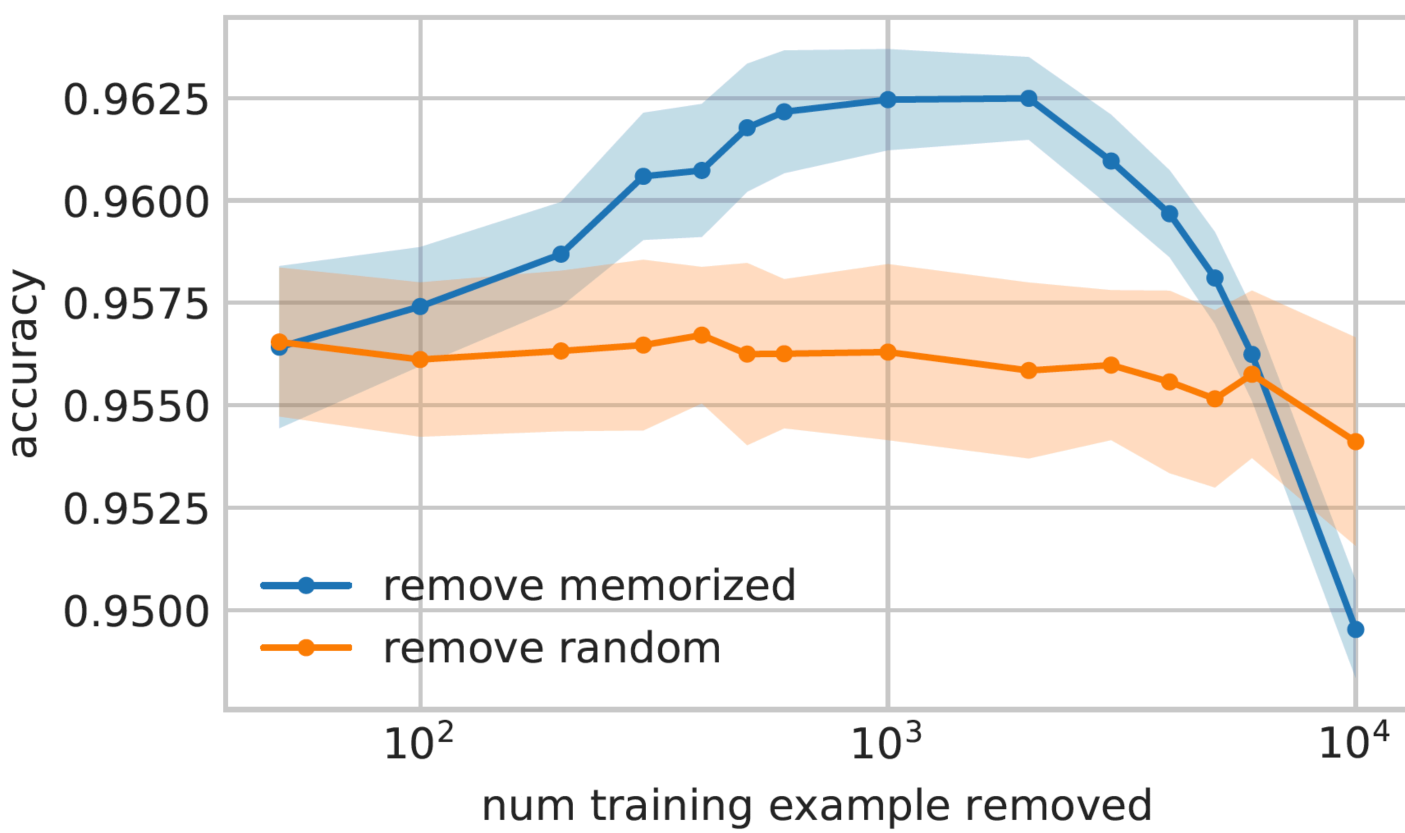
Ease of learning an instance in the training set is a good proxy for the probability that instance would be classified correctly were it held out from the training set.

ICML International Conference On Machine Learning

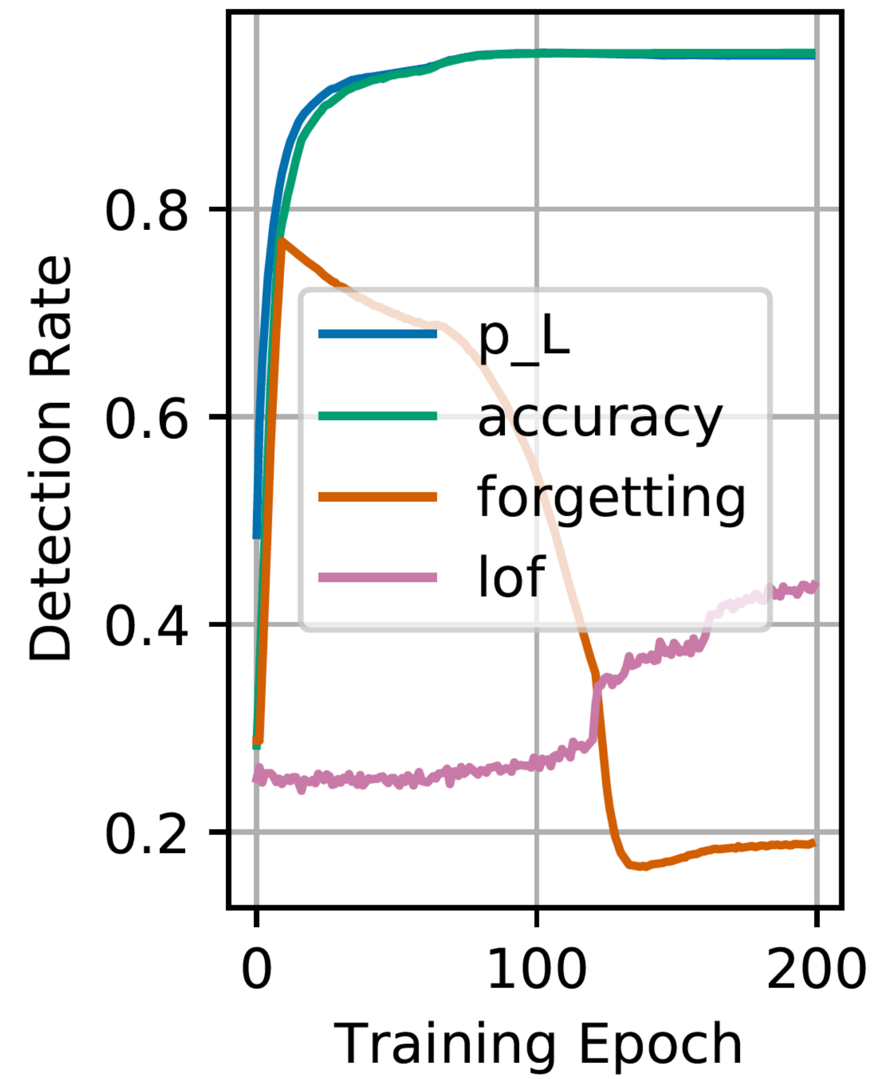
Application

ICML
International Conference
On Machine Learning

Application: Identify Outliers



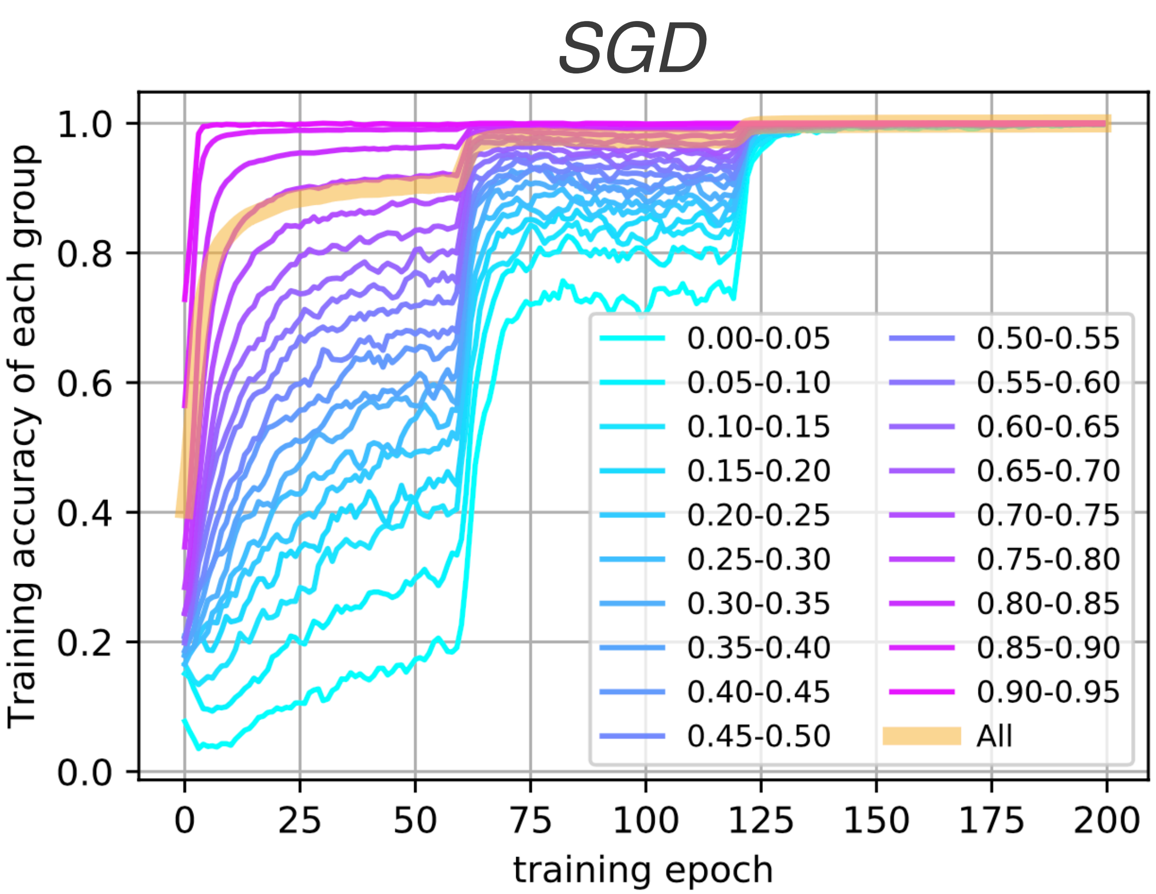
Model performance on SVHN when certain number of examples are removed from the training set.



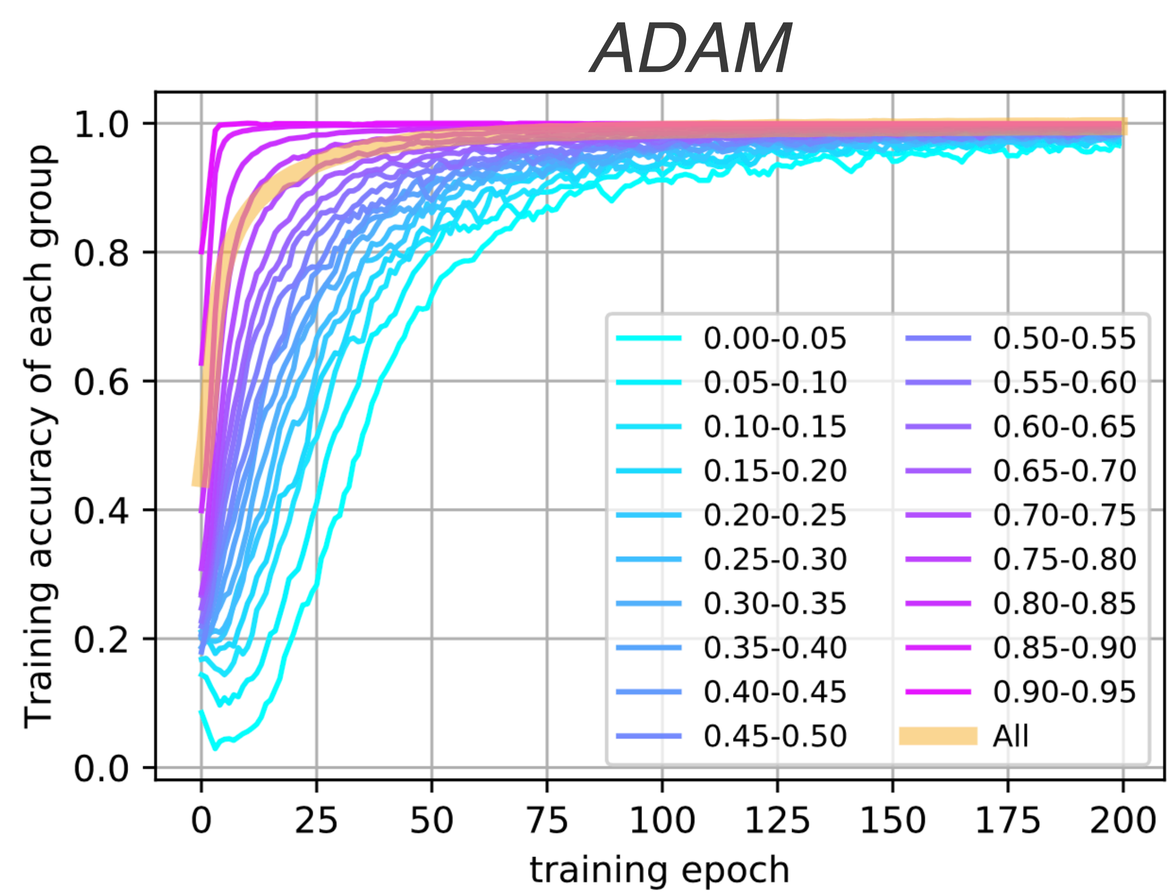
Detection rate of label-flipped outliers on CIFAR-10.

Application: Study the Behavior of Different Optimizers

We partition the CIFAR-10 training set into subsets by C-score. Then we record the learning curves—model accuracy over training epochs—for each set:



final test accuracy: 95.14%



final test accuracy: 92.47%

ICML
International Conference
On Machine Learning

Conclusion

Conclusion

- We introduce the C-score for individual instances in a data set
    - C-score: measure of how well an instance will generalize if it were held out of training

Conclusion

- We introduce the C-score for individual instances in a data set
- We compute empirical C-scores for all instances in CIFAR-10, CIFAR-100, MNIST, and ImageNet
    - Precomputed c-scores and algorithm code are available at https://pluskid.github.io/structural-regularity/

Conclusion

- We introduce the C-score for individual instances in a data set
- We compute empirical C-scores for all instances in CIFAR-10, CIFAR-100, MNIST, and ImageNet
- We use the C-scores to illustrate structural regularities in the data sets

Conclusion

- We introduce the C-score for individual instances in a data set
- We compute empirical C-scores for all instances in CIFAR-10, CIFAR-100, MNIST, and ImageNet
- We use the C-scores to illustrate structural regularities in the data sets
- We study computationally efficient proxies for the C-score
    - The amount of training required to learn an instance in the training set is a good predictor of generalization to that instance which is held out of training.

ICML
International Conference
On Machine Learning

Conclusion

- We introduce the C-score for individual instances in a data set
- We compute empirical C-scores for all instances in CIFAR-10, CIFAR-100, MNIST, and ImageNet
- We use the C-scores to illustrate structural regularities in the data sets
- We study computationally efficient proxies for the C-score
- We use the C-score and its proxy to
  - analyze the relative performance of ADAM and SGD with learning rate step downs
  - perform outlier detection and removal